

DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism

Year: 2021 | Citations: 318 | Authors: Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Zhou Zhao

Abstract

Singing voice synthesis (SVS) systems are built to synthesize high-quality and expressive singing voice, in which the acoustic model generates the acoustic features (e.g., mel-spectrogram) given a music score. Previous singing acoustic models adopt a simple loss (e.g., L1 and L2) or generative adversarial network (GAN) to reconstruct the acoustic features, while they suffer from over-smoothing and unstable training issues respectively, which hinder the naturalness of synthesized singing.

In this work, we propose DiffSinger, an acoustic model for SVS based on the diffusion probabilistic model. DiffSinger is a parameterized Markov chain that iteratively converts the noise into mel-spectrogram conditioned on the music score. By implicitly optimizing variational bound, DiffSinger can be stably trained and generate realistic outputs.

To further improve the voice quality and speed up inference, we introduce a shallow diffusion mechanism to make better use of the prior knowledge learned by the simple loss. Specifically, DiffSinger starts generation at a shallow step smaller than the total number of diffusion steps, according to the intersection of the diffusion trajectories of the ground-truth mel-spectrogram and the one predicted by a simple mel-spectrogram decoder. Besides, we propose boundary prediction methods to locate the intersection and determine the shallow step adaptively.

The evaluations conducted on a Chinese singing dataset demonstrate that DiffSinger outperforms state-of-the-art SVS work. Extensional experiments also prove the generalization of our methods on text-to-speech task (DiffSpeech). Audio samples: <https://diffsinger.github.io>. Codes: <https://github.com/MoonInTheRiver/DiffSinger>.